Kelly Hancox
CIS 678 Machine Learning
Programming Project #1

Experiments

Reading

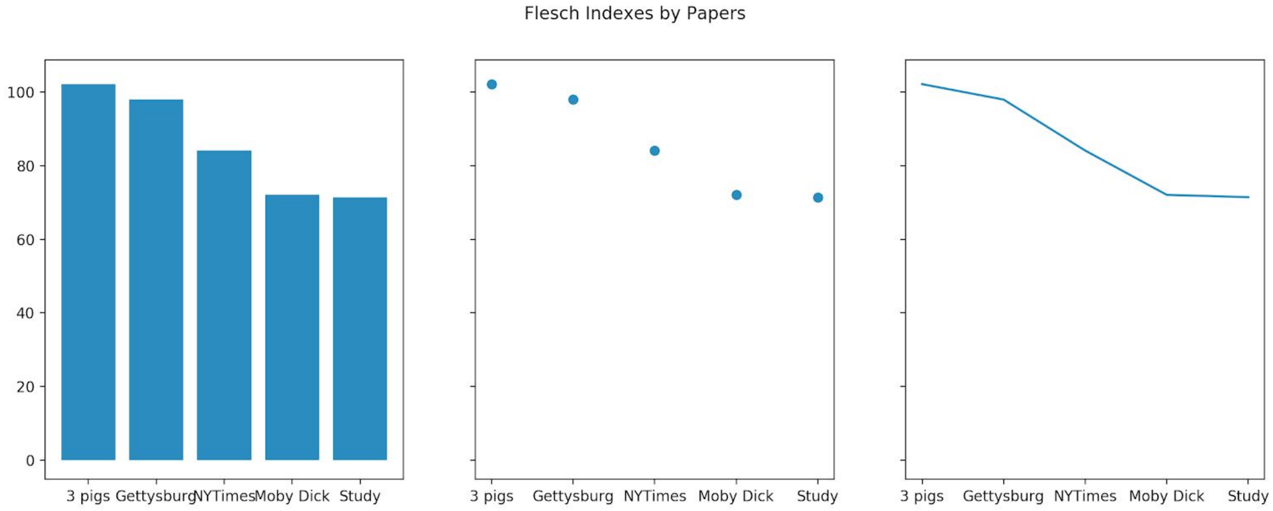
There were many ways to read in the document, either line by line, character by character, or trying to split the document by a delimiter. I decided to read it in character by character which worked for the most part, but then limited me when I tried to look for certain patterns.

Ellipsis

One of those patterns was the “. . .” ellipsis in the Gettysburg Address. I was trying to get each period encountered to add both a sentence and a word to the overall counts, but these counted for neither. I decided that in order to limit the amount of memory that my porram took that I would only look at the previous character and if it was not an alphabetic character, then I would not add to either count. This helped with the two additional periods, but still accounted for one extra sentence> Since it was a consistent rule throughout the other papers, I decided that this was okay.

Graphing

I first graphed the flesch indexes of 5 papers against one another using matplotlib and got this result:



This supported my assumptions that an easier paper would have a higher flesch index and a higher level study would have a lower flesch index. I added the 3 little pigs as a test document and a text file of a psychological study that I found on the internet. I used this as a practice to understand graphing with matplotlib.

Syllable Counts

The next data I wanted to gather was the amount of words that were different syllable lengths. To do this, I counted up the number of syllables and then restarted the count and added a count to this number of syllables within an array.

Anomalies

Capital Letters

Since I read the document in by character, I was checking if the characters would be alphabetic by putting the alphabet in two arrays, one for vowels and another for consonants. Something I overlooked in the first iteration was to include capital letters. I had surprisingly accurate data despite this oversight, but then changed this to change capital letters to lowercase letters later.

Punctuation

Another oversight after the ellipsis epiphany was that I had to check the previous character to determine the end of a word or not. Because of this, I had to alter my methods to also look for punctuation like a colon or semicolon or comma before whitespace.

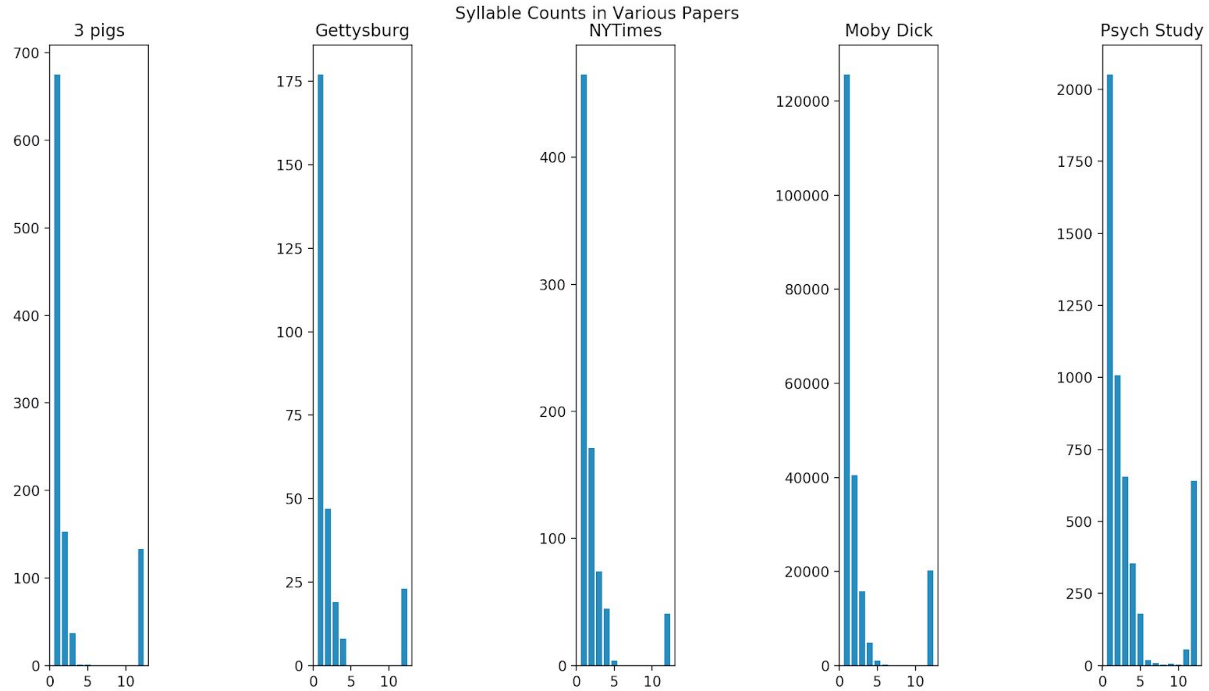
Syllable Count Data

When I was gathering data for syllable counts of words in documents, I was using the same system that I had used for counting the total amount of syllables in the document, where I was subtracting from the number of syllables in the document if I reached a whitespace or sentence end and checked the ‘e’. This did not quite work because I was subtracting from the amount of syllables after ending the word and I had to change that check to the beginning of that if statement.

An anomaly I encountered about this data was that I am seeing multiple words in every document at 12 syllables and I think that this data is incorrect, but I looked for a bug in this and did not find a place where

Conclusions

I found that, as expected, the more difficult texts also had words with higher numbers of syllables on average. This conclusion correlates with the Flesch Index formula, but the difference between the Psychological Study and Moby Dick was fascinating. The psychological study seemed to have many more words that had higher numbers of syllables, but its flesch index was not much higher than Moby Dick’s. Below the graphs are the raw data numbers.



3 Pigs [675, 153, 37, 1, 1, 0, 0, 0, 0, 0, 0, 133]

Gettysburg [177, 47, 19, 8, 0, 0, 0, 0, 0, 0, 0, 23]

NYTimes [465, 171, 74, 45, 4, 0, 0, 0, 0, 0, 0, 41]

Moby Dick [125654, 40446, 15800, 4864, 1086, 178, 16, 4, 0, 0, 95, 20250]

Psychological Study [2051, 1008, 655, 356, 180, 20, 8, 4, 6, 3, 56, 642]

Sources

https://matplotlib.org/3.1.1/gallery/subplots_axes_and_figures/subplots_demo.html